Topic Maps Extraction on Focused Web Pages By Clustering with Web Structure and Contents

Sa Mie Rar, Myo Kay Khaing
University of Computer Studies, Yangon
samierarlovely@gmail.com

Abstract

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce. Web mining is often associated with IR and IE.Web page clustering is one of the major preprocessing steps in web mining analysis.In this paper, clustering method is proposed to semi-automatically extract Topic Maps from a set of web pages. Firstly, the web pages are downloaded from the internet. To extract contents, remove stopwords and stem. And Second calculate the TF_IDF, content similarity, link similarity. Finally, calculating the Newman's method with the weighting based on the similarities by contents of web pages and types of links is applied to develop the potential clusters. Then the system generates the topic map by assuming the clusters as topics, the edges as associations, the web pages related to the topic as occurrences from the result of clustering. As the experimental result, more compact and denser cluster.